Schizophrenia Bulletin
https://doi.org/10.1093/schbul/sbac051

Natural Language Processing and Psychosis: On the Need for Comprehensive Psychometric Evaluation

Alex S. Cohen[*,1,2], Zachary Rodriguez[1,2], Kiara K. Warren[1], Tovah Cowan[1], Michael D. Masucci[1], Ole Edvard Granrud[1], Terje B. Holmlund[3], Chelsea Chandler[4,5], Peter W. Foltz[4,5], and Gregory P. Strauss[6]

[1]Louisiana State University, Department of Psychology, Baton Rouge, LA, USA; [2]Louisiana State University, Center for Computation and Technology, Baton Rouge, LA, USA; [3]University of Tromsø—The Arctic University of Norway, Tromso, Norway; [4]University of Colorado, Institute of Cognitive Science, Boulder, CO, USA; [5]University of Colorado, Department of Computer Science, Boulder, CO, USA; [6]University of Georgia, Department of Psychology, Athens, GA, USA

*To whom correspondence should be addressed; Louisiana State University, Department of Psychology, 210a Audubon Hall, Baton Rouge, LA, 70803, USA; tel: (225) 578-7017, Fax: (225) 578-4125, e-mail: acohen@lsu.edu

Background and Hypothesis: **Despite decades of "proof of concept" findings supporting the use of Natural Language Processing (NLP) in psychosis research, clinical implementation has been slow. One obstacle reflects the lack of comprehensive psychometric evaluation of these measures. There is overwhelming evidence that criterion and content validity can be achieved for many purposes, particularly using machine learning procedures. However, there has been very little evaluation of test-retest reliability, divergent validity (sufficient to address concerns of a "generalized deficit"), and potential biases from demographics and other individual differences.** ***Study Design:*** **This article highlights these concerns in development of an NLP measure for tracking clinically rated paranoia from video "selfies" recorded from smartphone devices. Patients with schizophrenia or bipolar disorder were recruited and tracked over a week-long epoch. A small NLP-based feature set from 499 language samples were modeled on clinically rated paranoia using regularized regression.** ***Study Results:*** **While test–retest reliability was high, criterion, and convergent/divergent validity were only achieved when considering moderating variables, notably whether a patient was away from home, around strangers, or alone at the time of the recording. Moreover, there were systematic racial and sex biases in the model, in part, reflecting whether patients submitted videos when they were away from home, around strangers, or alone.** ***Conclusions:*** **Advancing NLP measures for psychosis will require deliberate consideration of test-retest reliability, divergent validity, systematic biases and the potential role of moderators. In our example, a comprehensive psychometric evaluation revealed clear strengths and weaknesses that can be systematically addressed in future research.**

Key words: machine learning, reliability, validity, psychometrics, psychosis, paranoia, bias

Natural Language Processing (NLP) and Proof of Concept

Natural Language Processing (NLP) is a multidisciplinary pursuit that involves computational processing, analyzing, and quantifying various aspects of language. It has become indispensable to modern society. NLP has also been critical for understanding language, and how it relates to cognition, affect, and social functions. It is increasingly being used to "digitally phenotype" aspects of psychosis-spectrum disorders[1–3]; an endeavor that could reshape diagnosis, assessment, and treatment. From a pragmatic perspective, NLP allows for automation that can enhance traditional assessment by improving the efficiency, ecological validity, and accuracy of data collection, for example, by using data collected using mobile recording devices and social media platforms as individuals navigate their daily routines.[4,5] The use of "big" and high-dimensional data, collected on large, and demographically heterogeneous samples can also help address interpretive and practical constraints on traditional assessments.[6] NLP can facilitate systematic and repeated assessment (e.g., hourly, daily, weekly); important because "naturalistic" measurements can be less affected by learning and practice effects than traditional clinical measures.[7,8] In all, NLP can improve accuracy for objectifying relatively specific aspects of psychopathology.[9–11]

Clinicians have long used language to understand psychosis, and at least 5 decades ago began to use NLP to objectify and automate this process.[12,13] Since that time, NLP

© The Author(s) 2022. Published by Oxford University Press on behalf of the Maryland Psychiatric Research Center.
All rights reserved. For permissions, please email: journals.permissions@oup.com

has been used in a large and complicated literature spanning psychology, psychiatry, engineering, computational, and other clinical, and basic sciences.[1–3,14,15] NLP has been used to both measure and understand various aspects of psychosis, including present, and future DSM IV/5 diagnosis,[16,17] illness onset,[18] negative symptoms,[19–21] thought/language disorder,[22–24] social functioning,[25] hallucinations,[26] cognition,[27] paranoia,[13,25] substance use,[28] hope,[29] obsessive-compulsive symptoms,[30] and anhedonia.[31] It has been applied to a variety of media as well, such as online social media,[25] medical records,[32] standardized cognitive test responses,[33,34] clinical interviews,[18] autobiographical monologs,[31] and ambulatory audio/video recordings.[35,36]

Despite decades of "proof of concept" data supporting the use of NLP in psychosis research, there has been no implementation for psychiatric or psychological clinical care to our knowledge, nor has there been approval by governmental regulatory agencies for clinical trials, forensic use, or for service determination. While there is aggregate support for NLP to understand psychosis, there is limited support/replication for any specific NLP solution. In part, this reflects the reality that NLP involves a broad set of analytic methodologies focusing on potentially disparate aspects of language (e.g., syntactic, lexical, semantics) and reflecting different training corpuses[3] and different patient needs.[37] Moreover, strategies for evaluating NLP measures of psychosis have been relatively constrained, focusing primarily on a few aspects of validity with little attention to reliability, specificity, and bias. This article will a) discuss critical, but overlooked, aspects of psychometrics for evaluating NLP-based measures of psychosis, and b) highlight them by evaluating an NLP measure meant for tracking paranoia. In doing so, we hope to demonstrate a general psychometric approach for overcoming a major obstacle in implementing these measures.

Reliability, Validity and Bias

Traditionally, measures of psychiatric/psychological phenomenon are evaluated using reliability, and validity, the former being considered a prerequisite for the latter. Reliability concerns the consistency of a measure: across *time* (test–retest reliability), *individual items of a measure* (e.g., internal consistency), *informants* (e.g., inter-rater reliability), and *situations* (e.g., situational reliability). Validity concerns the measurement accuracy of its intended constructs; evaluated based on *putative structure* (e.g., structural validity) and potential *convergence with conceptually related* (e.g., convergent measure) and *unrelated* (e.g., divergent validity) constructs, and *clinically-relevant criterion* (e.g., concurrent, and predictive criterion validity). The precise aspects of reliability and validity, and their benchmarks, vary as a function of the nature of the measure and its application.[38] An NLP measure for monitoring change in disorganization using social media posts in European teenagers will differ in psychometric evaluation from a measure measuring alogia during clinical interviews in Asian adults.

There are several broad concerns with how NLP-based measures of psychosis have been evaluated thus far. First, reliability is rarely reported. Of the 206 peer-reviewed articles identified from an EBSCO search conducted on 10/24/2021 using terms "natural language processing" and "psychosis or schizo*" revealed 206 peer-reviewed entries, none presented reliability data. This is by no means a comprehensive literature search, but highlights the focus on validity over reliability. Insufficient attention to reliability, notably test-retest reliability has been identified as a major concern with network analysis,[39] taxometrics,[40] machine learning,[41] neuroscience[42] and mobile assessment[43,44] more generally. Elsewhere, we have evaluated test-retest reliability of computerized vocal and facial features and found it was often unacceptable without considering the influence of "moderating" variables, such as time of day, scope of assessment, and social factors .[45,46]

Second, validity is often evaluated with respect to sensitivity (e.g., criterion validity: convergence with a "gold standard" measure) with only superficial evaluation of specificity. The "generalized deficit" issue, concerning the false appearance of specificity due to more global group differences, has long been recognized as a potential confound in psychopathology assessment.[46–48] Language abnormalities can reflect a variety of cognitive, emotional, and psychological factors, many of which are nonspecific to schizophrenia (e.g., social impoverishment, chronic stress). NLP solutions often show impressive accuracy in differentiating patients from nonpatients, or symptomatic from asymptomatic patients. However, it is often not clear the degree to which this specificity is masked by nonspecific factors. Confounding this process is the loss of interpretability of original features when using machine learning solutions. In most NLP studies, features are engineered/optimized beyond recognition because knowing what was predicted when performance was good is sufficient. However, interpretability and "explainability", are increasingly being recognized as critical for the implementation of NLP technologies .[49]

Third, systematic influences on reliability and validity, such as from demographic, cultural, linguistic, and other individual differences, are often not considered. Systematic "biases" have long been a focus of psychometric evaluation,[50] and of NLP more generally.[51,52] Efforts to address them have been increasing within the last few years,[1] though to our knowledge, the potential effects on NLP measures of psychosis are poorly understood and rarely examined.[6,53] This is important for NLP to help address issues of systemic racism and inequality in psychosis assessment/treatment .[54,55]

There are challenges in applying traditional psychometrics to NLP. First, some aspects of reliability and validity don't readily apply to NLP. NLP solutions are

generally not meant to comprehensively capture constructs (i.e., latent phenomena), so evaluating the conceptual inter-relations of individual items/features through internal consistency, and construct and structural validity is often of limited use and considered unnecessary .[56] Moreover, the automated and objective nature of NLP obviates traditional inter-rater reliability, or at least, limits it to using different computers, datasets, software packages, and referential corpora if relevant. A greater challenge in evaluating NLP solutions involves the reality that language is highly dynamic within people, over time (e.g., min, hr, and days), and as a function of speaking task, environment, and other psychologically-relevant "moderators" (e.g., stress levels).[57,58] This can pose a challenge for establishing test-retest and situational reliability.

NLP Measure of Paranoia: An Example

Paranoia, defined in terms of self-relevant persecution, threat, or conspiracy from external agencies,[59] is a transdiagnostic and pernicious symptom of serious mental illnesses (SMI) that is typically measured using clinical ratings and self-report. Efforts to measure paranoia using NLP have been undertaken for over five decades, typically using lexical "text search" approaches using predefined dictionaries of word tokens (e.g., "good", "bad", "afraid"). Language samples typically involve autobiographical monologues and medical records from clinical settings.[13,32,60,61] Not surprisingly, symptoms of paranoia have been associated with relatively high negative affective and low affiliative word use. While promising in criterion and convergent validity, there has been limited evaluation of test–retest reliability, divergent validity/specificity, and systematic biases associated with these measures. Lexical expression, particularly negative affect, is likely a nonspecific measure of paranoia in that abnormalities are central to many types of psychopathology and symptoms (e.g., depression, anxiety). Moreover, they may differ in use by demographic factors. Clinically rated paranoia has been higher in Black than White samples[62] and likely reflects the interplay of a complex set of cultural, interpersonal, and professional influences.[63]

NLP Measure of Paranoia: NLP Feature Development/ Selection

NLP feature development and selection is a critical component of model development, and there are a variety of data and conceptually driven approaches used. Within paranoia research, studies have tended to use "out of box" solutions that may be nonspecific to paranoia; tapping negative affect more generally. Developing a small-feature set that aligns with social cognitive definitions of paranoia [64] may improve the psychometric characteristics, in particular, specificity to paranoia. In the next section, we evaluated several NLP measures for tracking relatively subtle fluctuations in paranoia from brief, topically-flexible speech samples procured from a smartphone while individuals navigated their daily routines. Patients with serious mental illness (SMI) provided "video selfies" over a week-long period. We employed an NLP procedure called sentiment analysis,[65] which is similar to lexical analysis approaches but takes into account valence shifters (e.g., negators, amplifiers) at a relatively fine-scale (phrases and complex word structures). Given that the operational definition of paranoia involves threat directed from the outside to the self, we focused on "self-other" references, defined as language that contained reference to both self and "non-self" using an NLP procedure that identifies subject-object dependencies within language. Finally, we examined the role of three moderators that putatively exaggerate, or at least concomitantly occur, with paranoia. Based on evidence that paranoia is dynamic as a function of being alone, around strangers, and experiencing threat,[66] we considered whether language was produced whether the individual was away from home, whether they were around strangers, and whether they were alone. Our focus on "self-other" speech and potential moderators was intended to improve specificity to paranoia beyond more global aspects of psychopathology.

We employed regularized regression to develop a model predicting clinically rated paranoia based on three NLP features, with consideration of three moderators (i.e., away from home, among strangers, and alone; see figure 1). Our NLP measure was evaluated in terms of a) test-retest reliability, b) convergence with clinical ratings of paranoia and self-report ratings of fear in the moment, divergence with measures of anxiety/depression, and c) bias in demographic factors.

NLP Measure of Paranoia: Methods

Participants:

Data was collected from 35 individuals with DSM-5[59] diagnoses of schizophrenia (SZ; n = 31) or bipolar disorder (BP; n = 4). Patients were primarily female (n = 10 men, 25 women) and White (23 White, 10 Black, 2 missing data) with an average age of 40.7 ± 11.5 and education of 13.7 ± 2.6 yr. Individuals with SZ were recruited from local community outpatient mental health centers. Clinical diagnosis was determined via the Structured Clinical Interview for DSM-5.[67] Participants provided written informed consent and received monetary compensation for their participation. This project was approved by the University of Georgia and Louisiana State University Institutional Review Boards and executed in accordance with internationally-recognized standards for the ethical conduct of human research .[45,68]

Data Collection:

Feature 1: Sentiment

Sentiment analysis objectifies affective tone of text based on keyword and phrase search, with scores ranging from -1 (extremely negative) to 0 (neutral) to 1 (extremely positive). This feature is the sentiment value for the entire language sample.

Feature 2: Self-Other Dependencies

Open Information Extraction identifies relational triples including a subject, object, and their relationship. For this project, we were interested in dependencies that included a "Self" and a "Other" in the subject-object pair. This feature is the number of Self-Other dependencies.

	Subject	*Relationship*	*Object*
Dependency 1	I	Am worried	Things
Dependency 2	The government	Harm	Me

Feature 3: Self-Other Sentiment

Sentiment analysis was conducted on the Self-Other dependencies. This feature is the most negative sentiment value of the independent Self-Other dependencies from the language sample.

Moderating Variables:

Three moderating variables were expected to exacerbate paranoia, and intensify the paranoia-related features

1. *Being Away from home*: Evaluated using the geolocation stamp while speaking. Values greater than 50 meters from home were counted as being outside the home.
2. *Being With Strangers*: Evaluated based on self-report just prior to speaking.
3. *Being Alone*: Evaluated based on self-report just prior to speaking.

Psychometric Evaluation Plan:	
Reliability	
Test-Retest	Relationship between NLP measures across testing sessions
Internal Consistency	Not applicable: NLP Measure not capturing latent phenomenon
Informants	Not applicable: NLP Measure not capturing latent phenomenon
Situational	Relationship between NLP measures across moderating variables
Validity	
Criterion	Relationships between NLP measures & Clinically Rated Paranoia
Convergent	Relationship between NLP measures &self-reported paranoia & fear
Divergent	Lack of relationship between NLP measures & depression/anxiety
Structural	Not applicable: NLP Measure not capturing latent phenomenon
Biases	
Demographic	Differential criterion validity due to ethnicity, sex, age and education

Fig. 1. Study rational and methods: description of the three key features and moderators developed and evaluated in this study, and the psychometric evaluation plan.

Clinical Measures

The Positive and Negative Symptom Scale (PANSS)[69] was used to measure psychiatric symptoms. The "Suspiciousness/Persecution" scale was used as our criterion, and the "Depression/Anxiety" factor score was used for divergent validity. Clinical ratings were made by staff trained to reliability standards (alpha >.8) on gold standard training tapes.

Mobile assessment

Participants were provided mobile phones preloaded with data collection software.[70] Surveys recorded current information such as whether the participant was "alone" at the time of the recording, and whether they were around "strangers". This information was coded as binary (i.e., yes, no) and not mutually exclusive. Self-reported in-the-moment "suspiciousness" and "fear" were also obtained (scale = 1–100), as was geolocation. Values greater than 50 meters of their home "pin" were deemed "away" from home.

Videos were recorded at the end of the surveys but were optional. Participants were instructed to record themselves while giving a step-by-step description of their past hour for 60 seconds. Participants were compensated $1 for each survey completed with no additional incentive for completing videos. Patients completed videos for an average of 23% (standard deviations = 20%) of the surveys. 499 videos were available for analysis.[45] Geolocation data were available for 368 videos, and self-report ratings were available for 412 videos. Approximately 45% of recordings were conducted away from home (164 of 368), 17% conducted with strangers (46 of 412), and 43% conducted alone (176 of 412). Missing data and distributions of surveys are in supplemental table 1.

Natural Language Processing

Speech was transcribed into text documents by trained research assistants; approximately 10% of which were reviewed by a supervisor. Using Python (v3.8) and Natural Language Toolkit library (NLTK),[71] documents were filtered by length (character length > 50), contractions were expanded, stop-words (e.g., the, an, a) were removed, and words not in NLTK English dictionary were removed. For each document, Stanford CoreNLP[72] annotators were used to identify parts of speech, lemmatize words, parse sentence-level dependencies, and extract open-domain relation triples (OpenIE). This yielded "dependencies", representing a subject, a relation, and the object of the relation. Sentiment analysis, conducted using the SentimentR package,[65] provides summary scores of the emotional valence of text using a bipolar scale from -1 (extremely negative) to 0 (neutral) to 1 (extremely positive). For each document, we extracted three NLP features to employ in our models: 1) The sentiment from the entire language sample (i.e., "*Sentiment*"), 2) the number of self-other dependencies in the entire language sample (i.e., "*Self-Other Dependencies*"), and 3) the minimum value from sentiment analysis of each individual self-other dependency (i.e., "*Self-Other Sentiment*"). The minimum value reflects the most negative of the Self-Other dependencies in that sample.

Patients produced, on average, 18.9 ± 14.3 dependencies, of which 2.7 ± 2.0 contained self and non-self in the subject and object. The average sentiment of these within-speech dependencies was slightly positive (0.2 ± 0.2; range = 10), and the minimum (0.0 ± 0.3; range = -1.1 to 0.7) and maximum (0.3 ± 0.3; range = -1.0 to 1.2) sentiment values from these dependencies (evaluated within an individual language sample) suggested there was notable range within a sample.

Analyses

First, we extracted and evaluated our three NLP features. Evaluation focused on convergence with our criterion (i.e., Clinically Rated Paranoia), test–retest reliability (using Intra-class Correlation Coefficients; ICC), and relationship to demographic variables. Second, we modeled Clinically Rated Paranoia from our three NLP features using ridge regression, which employed a five-fold cross-validation process (using an 80–20% training-test split to fit the original). We computed three models, each with seven terms: the three NLP features, one moderator (e.g., being home versus away during the recording), and three NLP by moderator interactions. Models varied based on the moderator used (i.e., three moderators, three models). An "Integrated NLP Measure" was computed using the average of the three fitted final model scores. We were unable to fit them together using additional regressions because of missing moderator data (e.g., missing GPS data). Third, the Integrated NLP measure was evaluated in a) test-retest reliability, b) convergence with mobile measures of suspiciousness and fear, c) divergence with clinical ratings of depression/anxiety, and d) potential bias in demographic variables. Analyses were conducted in R (R Core Team, 2017). We were unable to nest data within each patient because we were modeling "2nd order" variables (e.g., PANSS scores). Given that the number of videos varied by participant, reliability analyses (i.e., involving ICCs) were conducted on data averaged by day. This helped standardize data as a function of time—since recording times potentially varied across participants. All Variance Inflation Factor scores were below 2.50 suggesting multi-collinearity was not an issue. NLP data were standardized and trimmed (i.e., at 3.50 SDs).

NLP Measure of Paranoia: Results

Step 1: Feature Evaluation

Preliminary Statistics (Supplemental Table 2). A good range of clinically rated paranoia was observed in

the sample, with 17 patients showing no/questionable clinically rated paranoia (k = 283 videos) and 18 patients showing mild or greater clinically rated paranoia (k = 216 videos). A correlation matrix of dependent variables, features, and machine learning based scores is provided in supplemental table 2. Paranoia was inversely associated with the number of videos submitted at a trend level (r[33] = -0.30, P = .08). Patients with bipolar disorder versus schizophrenia were lower in clinically rated paranoia at a trend level (t[34] = 2.19, P = .07),

Test–Retest Reliability. Average ICC values of the three features were good to excellent across days (range of ICC's = 0.63–0.83). Single ICCs were lower (range of ICC's = 0.21–0.45). This suggests single time-point scores were a reliable measure of average scores, but were not themselves stable over time.

Criterion Validity (Supplemental Table 2). The NLP features were quite modest in their relations to Clinically Rated Paranoia (range of r's = -0.06 to 0.05).

Step 2: Modeling Clinically Rated Paranoia from NLP measures

Model Development (Table 1, Supplemental Table 5). NLP features showed relatively similar RMSE and MAE values across training and test cases, suggesting that overfitting was not a major concern. Final models explained 10%, 5%, and 9% of the variance for the "Away from Home", "Around Strangers" and "Alone" models respectively. An "Integrated NLP Measure" was computed as the average of available model scores.

Step 3: Integrated NLP Measures of Paranoia

Test–Retest Reliability (Table 2). The Integrated Measure showed average and single ICC values of 0.91 and 0.62, suggesting good to excellent test-retest reliability for both average and single time-point applications. Average ICC values for the independent models were generally good to excellent for all data, and for White, Black, and Male, and Female participants. Single ICC values varied from poor to good.

Convergent/Divergent Validity (Supplemental Table 1). As expected, the Integrated NLP Measure was associated with significantly increased Clinical Ratings of Paranoia (r[460] = 0.42, P <.001), Momentary Ratings of Suspiciousness (r[410] = 0.20, P < .001), and Momentary Ratings of Fear (r[410] = 0.21, P < .001). Also as expected, the Integrated NLP Model was not significantly correlated with clinical ratings of depression/anxiety (r[460] = 0.00, P = 1.00). Clinically Rated Paranoia was similarly associated with both Momentary Ratings of Suspiciousness (r[410] = 0.31, P < .001) and Fear (r[410] = 0.30, P < .001), and was not related to Clinically Rated Depression/Anxiety (r[497] = 0.00, P = 1.00).

Table 1. Model statistics forregularized regressions

	Training case (80%)				Testing case (20%)			
	K	RMSE	MAE	R^2	*K*	RMSE	MAE	R^2
Model: away from home								
5-Fold average	276	1.32	1.09	0.10	69	1.34	1.09	0.10
Final model	345	1.32	1.09	0.10	–	–	–	–
Model: being with strangers								
5-Fold average	330	1.39	1.21	0.05	82	1.39	1.22	0.04
Final model	412	1.398	1.20	0.05	–	–	–	–
Model: being alone								
5-Foldaverage	330	1.36	1.15	0.11	82	1.35	1.15	0.08
Final model	412	1.35	1.14	0.09	–	–	–	–

Notes: K, number of samples; RMSE, root mean square error; MAE, mean absolute error. Final model reflects all cases.

Demographic Biases (Figure 2; Supplemental Table 6). The Integrated NLP Measure showed differential prediction of Clinically Rated Paranoia for race and gender. Correlations for White (r[371] = 0.46, P < .001) and for female (r[316] = 0.40, P < .001) participants were much higher than for Black (r[63] = 0.00, P = 1.00) and male (r[142] = 0.18, P = .03) participants. Neither Age nor Education were significantly associated with the Integrated Model (r's = 0.01 and -0.07, P's > .11).

Clinically Rated Paranoia was significantly higher in Black compared to White participants (t[2, 472] = 9.11, P < .001, d = 1.49). The Integrated NLP Measure was significantly different in Black and White participants (t[2, 472] = 5.02, P < .001, d = 0.65), though the magnitude as approximately half as large. Both Clinically Rated Paranoia and the Integrated NLP Measure were significantly different between Men and Women (P's < .001, d's = 0.84 and 0.82 respectively). Both Age and Education were significantly associated with Clinically Rated Paranoia (r[472]'s = 0.19 and -0.29, P's < .001). The integrated NLP measure was significantly associated with age and education (r's = 0.19 and -0.10 respectively).

Moderators differed as a function of demographics. Black and White participants were notably different in whether they shared videos while Away from Home (14% versus 49% of total videos; Chi-square = 5.30, P = .02) or Around Strangers (0% versus 14% respectively; Chi-square = 11.38, P < .001). Men and Women were different at a significant/trend level in whether they shared videos while Away from Home (58% versus 38% of total videos; Chi-square = 9.01, P= .003) or Around Strangers (22% versus 6% respectively; Chi-square = 26, P = .02) or Alone (54% versus 37% respectively; Chi-square = 3.56, P = .06). Increasing age was associated with being alone (r[410] = 0.15, P < .001) and being home (r[366] = -0.26, P < .001). Education was not significantly associated with any of the moderators. There were no dramatic differences (i.e., > 10%) for participants as a group, but Black participants produced many more surveys when

away from home (35%) compared to video "selfies" when away from home (14%) (supplemental table 6).

Conclusions and Future Directions

While there is no "one size fits all" psychometric evaluation strategy for NLP measures of psychosis, there are common psychometric components appropriate for most. Test–retest reliability, divergent validity (i.e., addressing specificity), and demographic bias are three metrics that have received limited attention to date and are likely relevant regardless of the media evaluated (e.g., social media, traditional neuropsychology test, clinical interview), temporal scope of the assessment (e.g., single clinical interaction, repeated measurement), or clinical scope/purpose. Evaluating them will be critical for realizing NLP's potential for measuring psychosis and for adopting them for institutional use [reference to Marder, Hauglid, and Palaniyappan commentaries].

The Integrated NLP-based measure of paranoia examined in this study was reliable over a week-long epoch, showed good convergence with our criterion and conceptually related measures, and showed specificity/divergence with global measures of negative affect and psychopathology. It also showed higher criterion validity

Table 2. Intra-class correlation coefficient (ICC) values for the three NLP-based models of paranoia as a function of all data and of race and sex

		Model 1: away from home	Model 2: being with strangers	Model 3: being alone	Integrated NLP measure
All data[1]		0.77 (0.40)	0.72 (0.34)	0.81 (0.46)	0.91 (0.62)
Race[2]	White	0.93 (0.71)	0.71 (0.33)	0.71 (0.33)	0.74 (0.37)
	Black	0.86 (0.56)	0.88 (0.59)	0.91 (0.68)	0.89 (0.68)
Sex[2]	Male	0.45 (0.17)	0.62 (0.29)	0.80 (0.43)	0.73 (0.10)
	Female	0.87 (0.62)	0.58 (0.26)	0.61 (0.28)	0.74 (0.42)

Average and single (in parentheses) ICC values are presented.
[1]Data averaged by day for 7 days; ICC values reflect stability across days,
[2] Data averaged over 4 days, due to missing data.

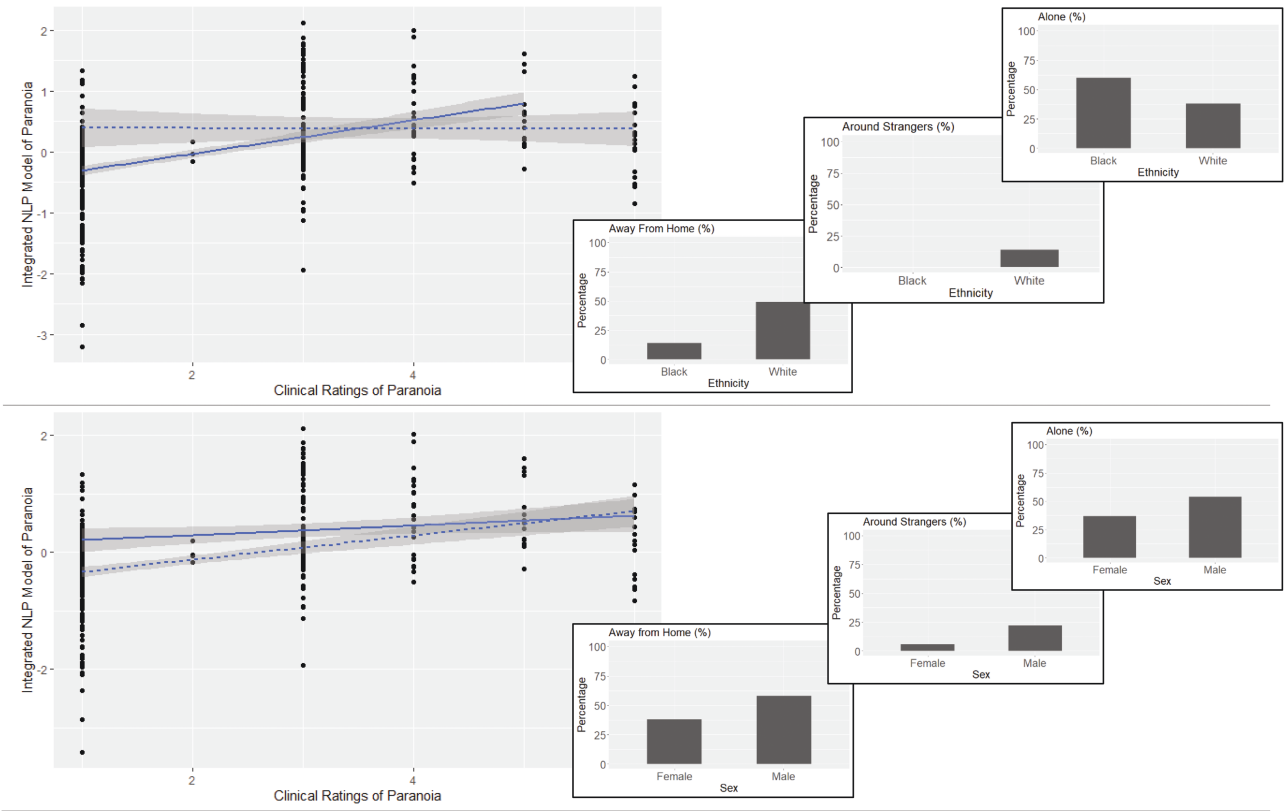

Fig. 2. Integrated NLP Solution as a function of race (top panel) and sex (bottom panel), with relative frequency of moderating variables (i.e., when videos were recorded). Solid line reflects white/male. Dotted line reflects black/female.

for White and Female participants than for Black and Male participants. This latter point underscores the need for comprehensive psychometric evaluations for NLP measures of psychosis since an evaluation strategy failing to consider demographic biases would have missed this critical issue. This study was limited in that the sample was modest in size, was demographically constrained, and video data were incomplete for many participants (due, in part, to a lack of compensation for data completion). The sample exclusively comprised people with SMI, and a community sample with a broader continuum of symptom severity would be important for future research. Moreover, reliability evaluation was limited to a single week in temporal scope; which is a limited epoch for understanding paranoia as a dynamic construct. Future research should include demographically diverse samples, in part, to power understanding of how demographic intersectionality (e.g., race, sex) affects measure psychometrics. Larger and more diverse samples are also critical for generalization.

While the present article was focused on general psychometric issues concerning NLP measures of psychosis, it is worth considering how one might address the sorts of biases found in our results. For Black participants, the bias appeared to reflect a lack of data across moderating conditions; participants were unlikely to provide videos when not at home or when around strangers. The reasons for this are not fully clear, though participants may have been reticent to provide data in these conditions due to concerns about personal safety, or stigma or negative evaluation from others. Paranoia, whether cultural, and/or mental illness in origin, likely did not attenuate these concerns. Outreach to potential participants as stakeholders will be critical in addressing these concerns, which could involve adapting data collection (e.g., using more subtle, less publicly noticeable procurement methods), providing additional participant support, training, compensation, or expanding the scope of assessment (e.g., to other situations where individuals experience paranoia but are comfortable providing videos). Men in this study showed lower criterion validity than women, and this did not seem to reflect a lack of data provision away from home or near strangers. The present sample size was not adequate for resolving demographic biases, inter-sectionality between demographic variables, and for generalizing our findings across diverse people and communities. While including a larger and more diverse sample than in this study is important, understanding the cultural differences in how paranoia manifests through language is also a critical concern highlighted by our data.

In closing, it is worth highlighting that language aspects were fairly dynamic over time. NLP data can be aggregated across language samples and across time and space/environment in various ways (e.g., both within and between testing sessions). This is important to consider for optimizing reliability, sensitivity, and specificity of a construct of interest[19,44,73] and is akin to "situational reliability", an aspect of psychometrics rarely examined in NLP-psychosis research. In the present data, NLP features varied as a function of moderating variables conceptually tied to paranoia. Our integrated model including these moderating variables showed improved reliability over most models/features not including them.

Realizing NLP's potential for measuring aspects of psychosis will require large amounts of complex data collected from geographically and culturally diverse groups,[74] and coordination between multidisciplinary and international groups.[75, 76] A comprehensive psychometrics strategy will be a critical part of this endeavor.

Supplementary Material

Supplementary material is available at https://academic.oup.com/schizophreniabulletin/.

Funding

This work was supported by NIH (National Institute of Health) (grant number R21 MH112925) to G.P.S. The authors have declared that there are no conflicts of interest in relation to the subject of this study.

Acknowledgment

We would like to acknowledge the efforts of the participants who contributed their data to the study, and the undergraduate students who helped process these data.
Competing interests: The authors declare that there are no competing interests.

References

1. Ratana R, Sharifzadeh H, Krishnan J, Pang S. A comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. *Front Psychiatry.* 2019;10:1–15. doi:10.3389/fpsyt.2019.00659
2. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: a natural language processing approach. *Schizophr Res.* 2020;226:158–166.
3. Holmlund TB, Fedechko TL, Elvevåg B, Cohen AS. Tracking language in real time in psychosis. In: *A Clinical Introduction to Psychosis.* Elsevier; 2020:663–685. doi:10.1016/B978-0-12-815012-2.00028-6
4. Cohen AS. Advancing ambulatory biobehavioral technologies beyond "proof of concept": introduction to the special section. *Psychol Assess.* 2019;31(3):277–284.
5. Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry connecting with technology. *JAMA Psychiatry* 2016;73(1):3–4.
6. Hitczenko K, Cowan HR, Goldrick M, Mittal VA. Racial and ethnic biases in computational approaches

to psychopathology. *Schizophr Bull.* 2022;48(2):285–288. doi:10.1093/schbul/sbab131.
7. Trull TJ, Ebner-Priemer U. The role of ambulatory assessment in psychological science. *Curr Dir Psychol Sci.* 2014;23(6):466–470.
8. Wright AGC, Zimmermann J. Applied ambulatory assessment: integrating idiographic and nomothetic principles of measurement. *Psychol Assess.* 2019;31(12):1467–1480.
9. Hsin H, Fromer M, Peterson B, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. *Npj Digit Med.* 2018;1(1):37.
10. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J.* 2015;38(3):218–226.
11. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. *Proc Natl Acad Sci.* 2018;201711978(27):E6106–E6115.
12. Maher B. The language of schizophrenia: a review and interpretation. *Br J Psychiatry.* 1972;120(554):3–17.
13. Colby KM. On the generality of PARRY, Colby's paranoia model. *Behav Brain Sci.* 1981;4:515–560.
14. Le Glaz A, Haralambous Y, Kim-Dufor D-H, et al. Machine learning and natural language processing in mental health: systematic review. *J Med Internet Res.* 2021;23(5):e15708.
15. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. *Curr Opin Psychiatry.* 2014;27(3):203–209.
16. Si D, Cheng SC, Xing R, Liu C, Wu HY. Scaling up prediction of psychosis by natural language processing. In: *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE; 2019:339–347. doi:10.1109/ICTAI.2019.00055
17. Voppel AE, de Boer JN, Brederoo SG, Schnack HG, Sommer IEC. Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Res.* 2021;304:114130.
18. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 2018;17(1):67–75.
19. Cohen AS, Schwartz EK, Le TP, et al. Digital phenotyping of negative symptoms: the relationship to clinician ratings. *Schizophr Bull.* 2021;47(1):44–53.
20. Cohen AS, Alpert M, Nienow TM, Dinzeo TJ, Docherty NM. Computerized measurement of negative symptoms in schizophrenia. *J Psychiatr Res.* 2008;42(10):827–836.
21. Cohen AS, Mitchell KR, Elvevåg B. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. *Schizophr Res.* 2014;159(2-3):533–538.
22. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1-3):304–316.
23. Elvevåg B, Foltz PW, Rosenstein M, DeLisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J Neurolinguistics* 2010;23(3):270–284.
24. Holshausen K, Harvey PD, Elvevåg B, Foltz PW, Bowie CR. Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex* 2014;55:88–96.
25. Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. In: *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality.* 2015:11–20.
26. Shiel L, Demjén Z, Bell V. Illusory social agents within and beyond voices: a computational linguistics analysis of the experience of psychosis. *Br J Clin Psychol.* 2022;61(2):349–363. doi:10.1101/2021.01.29.21250740
27. Holmlund TB, Chandler C, Foltz PW, et al. Applying speech technologies to assess verbal memory in patients with serious mental illness. *Npj Digit Med.* 2020;3(1):33.
28. Patel R, Colling C, Jyoti J, Jackson RG, Stewart RJ, Philip M. Illicit substance use in first episode psychosis (FEP): a natural language processing (NLP) electronic health record study. *Early Interv Psychiatry.* 2018;12(S1):99.
29. Bonfils KA, Luther L, Firmin RL, Lysaker PH, Minor KS, Salyers MP. Language and hope in schizophrenia-spectrum disorders. *Psychiatry Res.* 2016;245:8–14.
30. Chandran D, Robbins DA, Chang C-K, et al. Use of natural language processing to identify obsessive compulsive symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder. *Sci Rep.* 2019;9(1):14146.
31. Cohen AS, St-Hilaire A, Aakre JM, et al. Understanding anhedonia in schizophrenia through lexical analysis of natural speech. *Cogn Emot.* 2009;23(3):569–586.
32. Irving J, Patel R, Oliver D, … CC-S, 2021 undefined. Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. academic.oup.com. Accessed November 16, 2021. https://academic.oup.com/schizophreniabulletin/article-abstract/47/2/405/5918729
33. Holmlund TB, Cheng J, Foltz PW, Cohen AS, Elvevåg B. Updating verbal fluency analysis for the 21st century: applications for psychiatry. *Psychiatry Res.* 2019;273:767–769.
34. Chandler C, Holmlund TB, Foltz PW, et al. Extending the usefulness of the verbal memory test: the promise of machine learning. *Psychiatry Res.* 2021;297(4-7):113743. doi:10.1016/j.psychres.2021.113743.
35. Minor KS, Cohen AS. Affective reactivity of speech disturbances in schizotypy. *J Psychiatr Res.* 2010;44(2):99–105.
36. Abel DB, Minor KS. Social functioning in schizophrenia: comparing laboratory-based assessment with real-world measures. *J Psychiatr Res.* 2021;138:500–506.
37. Mohr DC, Weingardt KR, Reddy M, Schueller SM. Three problems with current digital mental health research... and three things we can do about them. *Psychiatr Serv.* 2017;68(5):427–429.
38. Kane MT. An argument-based approach to validity. *Psychol Bull.* 1992;112(3):527–535.
39. Forbes MK, Wright AGC, Markon KE, Krueger RF. Evidence that psychopathology symptom networks have limited replicability. *J Abnorm Psychol.* 2017;126(7):969–988.
40. Watson D. Investigating the construct validity of the dissociative taxon: stability analyses of normal and pathological dissociation. *J Abnorm Psychol.* 2003;112(2):298.
41. Hajcak G, Meyer A, Kotov R. Psychometrics and the neuroscience of individual differences: internal consistency limits between-subjects effects. *J Abnorm Psychol.* 2017;126(6):823–834.
42. Elliott ML, Knodt AR, Ireland D, et al. What is the test–retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol Sci.* 2020;31(7):792–806.

43. Cohen AS, Schwartz E, Le T, et al. Validating digital phenotyping technologies for clinical use: the critical importance of "resolution". *World Psychiatry* 2020;19(1):114–115.
44. Cohen AS, Cox CR, Tucker RP, et al. Validating biobehavioral technologies for use in clinical psychiatry. *Front Psychiatry.* 2021;12. doi:10.3389/fpsyt.2021.503323.
45. Cohen AS, Cowan T, Le TP, et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr Res.* 2020;220:141–146.
46. Chapman L, Bulletin JC-P. 1973. Problems in the measurement of cognitive deficits. *psycnet.apa.org*. 1973;79(6):380–385. Accessed November 20, 2021. https://psycnet.apa.org/record/1973-31688-001
47. Green MF, Horan WP, Sugar CA. Has the generalized deficit become the generalized criticism? *Schizophr Bull.* 2013;39(2):257–262.
48. Gold JM, Dickinson D. "Generalized Cognitive Deficit" in schizophrenia: overused or underappreciated? *Schizophr Bull.* 2013;39(2):263–265.
49. Chandler C, Holmlund TB, Folt PW, Cohen AS, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull.* 2020;46(1):11–14. doi:10.1093/schbul/sbz105.
50. Cole NS. Bias in testing. *Am Psychol.* 1981;36(10):1067–1077.
51. Leavy S. Gender bias in artificial intelligence. In: *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering.* ACM; 2018:14–16. doi:10.1145/3195570.3195580
52. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv.* 2021;54(6):1–35.
53. Schwartz EK, Docherty NM, Najolia GM, Cohen AS. Exploring the racial diagnostic bias of schizophrenia using behavioral and clinical-based measures. *J Abnorm Psychol.* 2019;128(3):263–271.
54. Minsky S, Vega W, Miskimen T, Gara M, Escobar J. Diagnostic patterns in Latino, African American, and European American psychiatric patients. *Arch Gen Psychiatry.* 2003;60(6):637–644.
55. Olbert CM, Nagendra A, Buck B. Meta-analysis of black vs. white racial disparity in schizophrenia diagnosis in the United States: do structured assessments attenuate racial disparities? *J Abnorm Psychol.* 2018;127(1):104–115.
56. Chapman BP, Weiss A, Duberstein PR. Statistical learning theory for high dimensional prediction: application to criterion-keyed scale development. *Psychol Methods.* 2016;21(4):603–620.
57. Wright AGC, Hopwood CJ. Advancing the assessment of dynamic psychological processes. *Assessment* 2016;23(4):399–403.
58. Cohen AS, Fedechko TL, Schwartz EK, et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *J Abnorm Psychol.* 2019;128(2):97–105.
59. American Psychiatric Association, D. S., & American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5 (Vol. 5)*. Washington, DC: American psychiatric association.
60. Oxman TE, Rosenberg SD, Tucker GJ. The language of paranoia. *Am J Psychiatry.* 1982;139(3):275–282. doi:10.1176/ajp.139.3.275.
61. Finn ES, Corlett PR, Chen G, Bandettini PA, Constable RT. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nat Commun.* 2018;9(1):2043.
62. Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. *World J Psychiatry.* 2014;4(*4*):133. Accessed March 21, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4274585/
63. Whaley AL. Ethnicity/race, paranoia, and psychiatric diagnoses: clinician bias versus sociocultural differences. *J Psychopathol Behav Assess.* 1997;19(1):1–20. Accessed March 21, 2018. https://link.springer.com/article/10.1007/BF02263226
64. Green MJ, Phillips ML. Social threat perception and the evolution of paranoia. *Neurosci Biobehav Rev*. 2004;28(3):333–342. doi:10.1016/j.neubiorev.2004.03.006
65. Rinker T. Package "sentimentr." Published online 2021. Accessed February 14, 2022. https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf
66. Fett AKJ, Hanssen E, Eemers M, Peters E, Shergill SS. Social isolation and psychosis: an investigation of social interactions and paranoia in daily life. *Eur Arch Psychiatry Clin Neurosci.* 2022;272(1):119–127. doi:10.1007/s00406-021-01278-4
67. First M, Williams J, Karg R, Spitzer R. *User's guide for the SCID-5-CV structured clinical interview for DSM-5® disorders: clinical version.*; 2016. Accessed December 8, 2021. https://psycnet.apa.org/record/2016-15667-000
68. Raugh IM, James SH, Gonzalez CM, et al. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. *Schizophr Bull.* 2020;46(6):1596–1607.
69. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276. Accessed August 30, 2018. https://academic.oup.com/schizophreniabulletin/article-abstract/13/2/261/1919795
70. mEMA. https://ilumivu.com/solutions/ecological-momentary-assessment-app/
71. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* In Steele J, ed. O'Reilly Media, Inc.; 2009.
72. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, Mcclosky D. The Stanford CoreNLP Natural Language Processing Toolkit. *In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.* 2014:55–60. Accessed February 14, 2022. https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf
73. Cohen AS, Schwartz E, Le TP, Fedechko T, Kirkpatrick B, Strauss GP. Using biobehavioral technologies to effectively advance research on negative symptoms. *World Psychiatry* 2019;18(1):103–104.
74. Stewart R, Velupillai S. Applied natural language processing in mental health big data. *Neuropsychopharmacology* 2021;46(1):252–253.
75. Palaniyappan L, Alonso-Sanchez M, MacWhinney B. Is collaborative open science possible with speech data in psychiatric disorders? *Schizophr Bull.* 2022;48(5):963–966.
76. Hauglid, M. What's the noise? Interpreting algorithmic interpretation of human speech as a legal and ethical challenge. *Schizophr Bull.* 2022;48(5):960–962.